AFRL-AFOSR-VA-TR-2015-0311

Lossy Information Exchange and Instantaneous Communication

Lizhong Zheng
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

09/17/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
10-09-2015	Final Report	July 2012 - July 2015

4. TITLE AND SUBTITLE

Lossy Information Exchange and Instantaneous Communications

5a. CONTRACT NUMBER FA9550-11-1-0168

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Lizhong Zheng

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT
77 Mass Ave.
Cambridge, MA02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A - Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this project, we studied a new theoretical framework for information transmission with the focus on extracting only a part of the information. This formulation is particularly useful when we apply information theory to data analysis problems, where the goal is different from the full and reliable information recovery needed in the classical communications problems. We developed a geometric structure for the space of probability distributions, and a new method to decompose the information carried by the observed data based on that. This formulation gives a general setup to understand dimension reduction as lossy information processing procedures and gives a new operational meaning of information metrics, in the context of data analytics. We can quantitatively describe the information efficiency, computation complexity, and sample complexity to learn the model in one picture. We also make connections to the existing results on dimensional reduction. Based on this framework, we developed new algorithms for dimension reduction and non-linear feature selection. We proved a number of optimality results for these new constructions, and

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
U	U	U	SAR		Lizhong Zheng
					19b. TELEPHONE NUMBER *(Include area code)* 617-452-2941

Reset

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Final Report on AFOSR Project "Lossy Information Exchange and Instantaneous Communications", FA9550-11-1-0168

Lizhong Zheng, September 2015

The goal of this projecct was to develop a new framework to study information exchanges with constraints in the delays and computational complexity, from an information theoretic perspective.

The project is based on a key difference between digital communication problems and statistical inference problems, namely, in communications, we often have the goal of delivering the information as an entirty, such that every bit of the information is reliably conveyed; whereas in inference problems, it is often the case that we are only interested in a special feature in the data, and do not wish to reconstruct the entire observed data after processing. This is a fundamental difference, as in the later cases, one has to distinguish between different components in an observation, make sure the relevant/important information remain intact after processing and transmission, and often discard the rest of the information. This notion of lossy processing, or information discipation in the processing, is a concept that is lacking the conventional information theoretic analysis.

We have, in the first a few years of this project, developed a new framework which we call ``linear information coupling". In this setup, we think any informtion exchange as a variation of the posterior distribution of the information-carrying message. While the variation of the distribution is often a very high-dimensional object, we define an ortho-normal basis in the space of probability distributions. We choose this basis in correspondence to the SVD spectrum of the observation model, so that we can decompose the calculation of the posterior distributions into computing a sequence of scores. These scores are ordered and labeled by the information contents they contain. With this general structure, we can, instead of computing the entire posterior distribution of the message, compute only a few, say, k, scores from the top of the list, knowing that these score values contains the maximum amount of useful information among all k-dimensional statstics that can be extracted from the observations. We call these scores "efficient statistics". They are not sufficient statistics in the conventional sense, as they do not contain all the information in the entire observation, but are the most informative functions one can compute, given the computation complexity that one can afford.

The idea of information coupling has been applied to a variety of problems. This includes new techniques that can be used for the traditional network commnication problems, as well as inference problems. Some of these results are reported in the following publications.

Shao-Lun Huang, Anuran Makur, Fabian Kozynski, Lizhong Zheng: Efficient Statistics: extraing Information Form IID Observations. Allerton Conference 2014

Shao-Lun Huang, Lizhong Zheng: The Linear Information Coupling Problems. CoRR abs/1406.2834

Shao-Lun Huang, Changho Suh, Lizhong Zheng: Euclidean information theory of networks. Accepted IEEE Trans. Info. Theory, August, 2015.

Near the end of this project, we have made another important finding. In all of our previous works, we assumed that the noisy observation channel model is precisely known. This is, however, a problematic assumption in reality. With the help of the SVD structure, we came up with an algorithm, which is a generalization of the Alternating Conditional Expectation (ACE) algorithm, which efficiently estimate the top singular vectors of the DTM, and use them as the score functions, or efficient statistics. This put our result in the general category of dimension reduction/non-linear feature selection problems.

Theoretically, the reason for this simple algorithm to work is related to the concept of Renyi maximal correlation. In 1959, Renyi proposed a measure of the dependence between two random variables, by finding a pair of functions, one for each of the random variables, which are highly correlated. In our application, one variable corresponds to the user, the other to the service. Finding the Renyi correlation in this problem can thus be understood as finding a particular aspect of the services that can be used to distinguish the users, i.e. to answer the question `which user is more likely to use a certain portfolio of services?' Our algorithm is in fact a very efficient way, not only to compute the maximal Renyi correlation, but also to find the best choices of 200 pairs of functions. As we distill everything we know about a user into a 200 dimensional signature, we inevitably discard some knowledge about this user; but with our algorithm, it is guaranteed that these are the 200 values that carry the most amount of useful information, for the purpose of predicting their service preferences.

The conceptual step made in this algorithm is that instead of estimating the complete statistical model of the observation channel, we can instead only estimate one “mode” of it, corresponding to the maximal Renyi correlation, and the singular vector of the DTM with the leading singular value. We show that this requires a significantly smaller number of training samples, which is the critical issue for most Big Data problems. On the other hand, the resulting statistics are also information theoretically optimal, in the sense that they carry the largest amount of information, for the given number of statistics. It is thus quite surprising that our result gives the optimal tradeoff between three objectives: the inference performance, the number of statistics used, and the sample complexity to learn these score functions.

There are some further advantages of our algorithm. The most useful one in practice is the generality of the algorithm. Instead of requiring specific forms of data, and restricting to specific type of score functions, such as linear functions over real-valued data in PCA, our algorithm is much more general, and can be used to process information with different types of data, discrete or continuous valued. The score functions selected from this approach are in general non-linear, which shed lights to the difficult problem of non-linear feature selection. We can also put additional constraints for specific types of score functions in our formulation. For example, when we apply to real valued data and insist the score functions to be linear, we indeed recover the PCA solution as a special case. Thus, our algorithm can be used in process and integrate data from different applications, to jointly improve the performance in inference.

At the very end of this project, we just started reporting our results in publications. The first one of them is

Anuran Makur, Shao-Lun Huang, Fabian Kozynski, Lizhong Zheng, “An Efficient Algorithm for Information Decomposition and Extraction” Allerton conference, Oct. 2015

We have applied this algorithm in a number of realistic problems, including the Netflix problem, a community detection problem with Facebook connection graphs, and a detection problem with audio signals.

1.

1. Report Type

Final Report

Primary Contact E-mail

Contact email if there is a problem with the report.

lizhong@mit.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

617-452-2941

Organization / Institution name

MIT

Grant/Contract Title

The full title of the funded effort.

Lossy Information Exchange and Instantaneous Communications

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-11-1-0168

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Lizhong Zheng

Program Manager

The AFOSR Program Manager currently assigned to the award

James Lawton

Reporting Period Start Date

07/01/2012

Reporting Period End Date

07/01/2015

Abstract

In this project, we studied a new theoretical framework for information transmission with the focus on extracting only a part of the information. This formulation is particularly useful when we apply information theory to data analysis problems, where the goal is different from the full and reliable information recovery needed in the classical communications problems. We developed a geometric structure for the space of probability distributions, and a new method to decompose the information carried by the observed data based on that. This formulation gives a general setup to understand dimension reduction as lossy information processing procedures and gives a new operational meaning of information metrics, in the context of data analytics. We can quantitatively describe the information efficiency, computation complexity, and sample complexity to learn the model in one picture. We also make connections to the existing results on dimensional reduction. Based on this framework, we developed new algorithms for dimension reduction and non-linear feature selection. We proved a number of optimality results for these new constructions, and applied the algorithm in real data analysis tasks.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF
The maximum file size for an SF298 is 50MB.

AFD-070820-035.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

Final Report on AFOSR Project 2015.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Changes in research objectives (if any):

Change in AFOSR Program Manager, if any:

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Sep 10, 2015 15:12:08 Success: Email Sent to: lizhong@mit.edu